Using facial feedback to enhance turn-taking in a multimodal dialogue system

To appear in the Proceedings of HCII-05

Using Facial Feedback to Enhance Turn-Taking in a Multimodal Dialogue System

Michael White, Mary Ellen Foster, Jon Oberlander, Ash Brown

School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, United Kingdom
{Michael.White, M.E.Foster, J.Oberlander}@ed.ac.uk, AshBrown@stanfordalumni.org

Abstract

We describe the results of an experiment investigating whether an avatar's facial feedback can enhance turn-taking, undertaken as part of a usability study of a preliminary version of the COMIC multimodal dialogue system. The study focused on the phase of the interaction where the avatar embodies a virtual sales agent that guides the user through a range of possible tiling options for his or her newly redesigned bathroom. Our experiment employed a between-subjects design, where subjects used the system in one of two face conditions: (1) the "expressive" condition, where lip sync, blinking, facial expressions, gaze shifting and head turning were enabled; or (2) the "zombie" condition, where only lip sync was enabled. The results of the study were mixed, with some positive results on improving the interaction quality, but some unexpected negative results on task success and ease. On the positive side, the responses to our questionnaire indicated that the avatar's thinking expression helped to convey that the system was busy processing input—confirming Edlund and Nordstrand's (2002) finding—and that the facial expressions mitigated the system's perceived sluggishness in responding verbally. However, after examining the videos of the interactions, we concluded that the avatar's facial feedback—though helpful with some users—was unlikely to make up for the unnaturalness of the system's half-duplex interaction on its own, and thus should be used together with explicit signals such as busy cursors. We did also find that the subjects in the expressive condition looked back at the avatar significantly more often than those in the zombie condition—confirming the results of Sidner et al. (2004)—but it was unclear whether this had any impact on turn-taking. Interestingly, recent research by de Ruiter (2005) revealed no systematic relationship between other-gaze and turn-taking in human-human dialogues involving relevant external visual representations, so in retrospect the absence of any significant impact of the avatar's looking behaviour on turn-taking should perhaps be expected. With task success and ease, we were surprised to find that the subjects in the zombie condition scored significantly higher on several of our objective and perceived measures. One reason for the negative impact of the expressive face on task success and ease may have been that the expressive face distracted subjects from the task. Another possibility is that the expressive face raised users' expectations of the system's abilities, thereby encouraging subjects to use voice input rather than the mouse, which was generally a less successful strategy. We plan to investigate further with the final version of the system.

1 Introduction

In this paper, we describe the results of an experiment investigating whether an avatar's facial feedback can enhance turn-taking, undertaken as part of a usability study of a preliminary version of the COMIC multimodal dialogue system (cf. den Os & Boves, 2004; Foster & White, 2004; White & Foster, 2004). The study focused on the phase of the interaction where the avatar embodies a virtual sales agent that guides the user through a range of possible tiling options for his or her newly redesigned bathroom.

As noted in Cassell et al.'s (2001) review, early work on avatars (or embodied conversational agents) showed that users often prefer interfaces with human faces to equivalent interfaces without an embodied agent, finding them more engaging or entertaining. More recently, Nakano et al. (2003) and Sidner et al. (2004) have shown that avatars that shift their looking from the user to the objects under discussion and back can influence how much attention a user pays to the face. However, there has been relatively little success so far in showing that the human faces can actually improve task performance or interaction quality. For this reason, we decided to investigate whether the COMIC avatar could improve usability, and in particular, whether its facial expressions and looking behaviour could make turn-taking somewhat more intuitive in a strict half-duplex interaction.

In our experiment, we hypothesized that the avatar's thinking expression displayed at the end of the user's turn would help to convey that the system was busy processing input, and that its subsequent nods, smiles or confused expressions would provide an early visual indication of the system's success in processing user input. We also hoped to show that the avatar's looking behaviour would improve conversational efficiency by helping to signal to the user when the system had come to the end of its turn, as suggested in (Cassell & Thorisson, 1999), and in line with Kendon's (1967) observation that the close of a turn may be signalled to some extent by the speaker looking more at the listener.

Our experiment employed a between-subjects design, where a total of 37 subjects used the system in one of two face conditions: (1) the "expressive" condition, where lip sync, blinking, facial expressions, and head turning were enabled; or (2) the "zombie" condition, where only lip sync was enabled. The results of the study were mixed, with some positive results on improving the interaction quality—in particular, that the avatar's thinking expression helped to convey that the system was busy processing input, and that the facial expressions mitigated the system's perceived sluggishness in responding verbally—but some unexpected negative results on task success and ease.

In the next section, we describe the capabilities of the system and sketch the user interaction. The body of the paper reports on the details of our experiment and the results. We conclude with a discussion of the implications of the results for multimodal system design, and outline the ways in which we have taken these results on board in developing the final version of the system.

2 Dialogue Capabilities

Users interact with the COMIC system in three main phases. In the first two phases, users enter the dimensions of their bathroom (with pen/mouse and voice commands) and choose a layout for the sanitary ware. In the third phase, studied here, the system guides the user through a selection of the available tiling designs. These designs consist of coherent sets of tiles, referred to internally as tilesets. In the display, the current tileset is shown in the user's bathroom, and along the bottom there are up to 5 thumbnails of other tilesets that the user may choose to look at. In addition to selecting thumbnails with the pen/mouse, the user may also ask to see designs in a certain style or with certain colours, or to see a 3-D tour. The system also suggests these options.

A screenshot of the display is shown in Figure 1 below. Two views of the face follow in Figure 2; on the left, the face is in a neutral position, while on the right, the eyes and head are turned to look at the bathroom display, and the eyebrows are raised in sync with word-level emphasis in speech.

In our study, the subjects sat in front of a table which had both the bathroom and avatar screens on it, in addition to a mouse and speakers. The screens were side-by-side and angled towards each other. Part of an automatically produced transcript of the dialogue with one of the subjects appears below in Table 1. Note that system phrases such as "this design" are accompanied by synthetic pointing gestures at the referenced thumbnail.

There were several known shortcomings in this version of the system, which we expected to have an impact on usability:

- Turn-taking was subject to a strict half-duplex protocol. The input channels were opened only after the system was finished producing its output; the user could not "barge in", with either pen or speech, while the system was speaking. There was no explicit indication of when the system was actually listening, as we were interested in seeing the extent to which the face could provide this information.
- The manufacturer and series names were displayed on the screen along with the tiles, but none of those were included in the ASR language model and therefore could not be recognized if the user attempted to choose a design via speech.
- There were delays in every module of the system, which compounded to make the interaction feel sluggish overall.

A half-duplex protocol was chosen for the COMIC system in part because of limitations in the public domain ASR module and lack of echo cancellation in the platform, and in part to avoid the complexities of incremental processing that would need to be confronted in every module of the system in order to handle full-duplex interaction.

Figure 1: Bathroom display

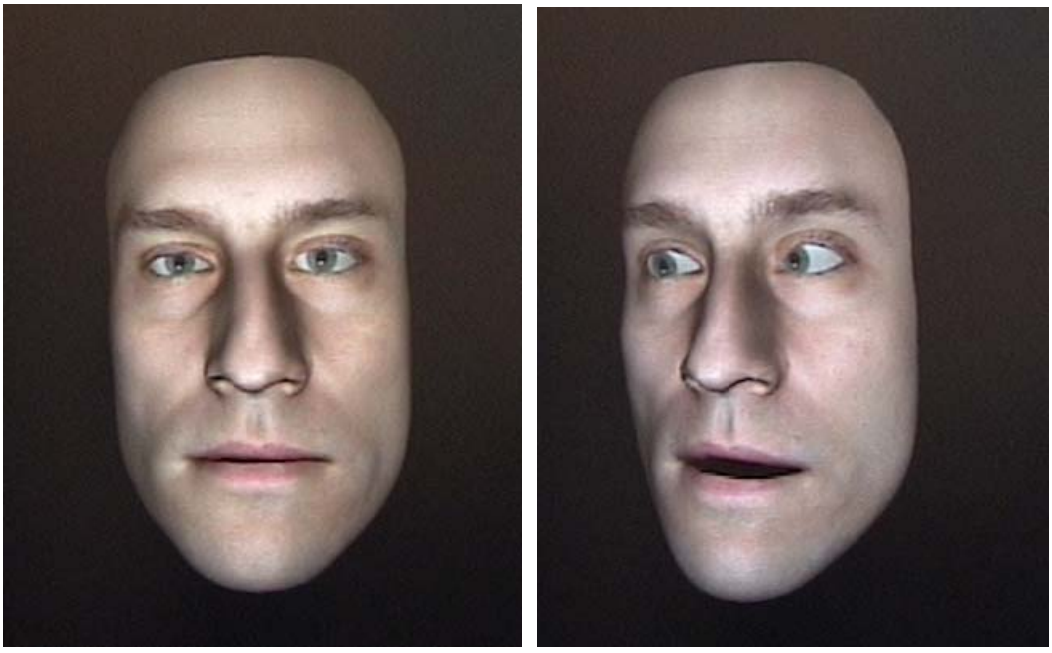

Figure 2: Avatar with neutral expression (left) and looking at the bathroom display (right)

Table 1: Sample dialogue fragment

```
SYSTEM: We are going to choose the tiles that you can use in your bathroom.
SYSTEM: We have a range of designs for you to look at, which show a variety of styles, colours,
and decorative ideas.
SYSTEM: Each design includes a small number of tiles that have been carefully selected to go
together well.
SYSTEM: I can help you find a design that you like, and then you can see a three dimensional tour
of the room.
SYSTEM: Are you ready?

USER: "yes"

SYSTEM: All right.
SYSTEM: This design is classic.
SYSTEM: The tiles are from the Alt Mettlach collection by Villeroy-and-Boch.
SYSTEM: This is one of our most popular collections.
SYSTEM: There are floral motifs and geometric shapes on the decorative tiles.
SYSTEM: The tiles have an off white, terracotta and black colour scheme.
SYSTEM: This design is modern, while this one is family.
SYSTEM: This design is modern, while this one is country.
SYSTEM: Please choose one of these designs to look at next.

USER: (No speech) [Choose tileset]

SYSTEM: Sure.
SYSTEM: As I mentioned earlier, this design is family.
SYSTEM: It's based on the Lollipop collection by Agrob Buchtal.
SYSTEM: There are floral motifs and artwork on the decorative tiles.
SYSTEM: The lollipops are always a treat for the kids.
SYSTEM: It features blue and green in the colour scheme.
SYSTEM: Would you like to see more family designs?

USER: "yes"

SYSTEM: All right. …
```

3 Experimental Design

As mentioned in the introduction, our experiment employed a between-subjects design, where subjects used the system in either the "expressive" or "zombie" condition. A total of 37 subjects participated in the experiment, split almost evenly (19/18) between the two conditions. Of the 37 subjects, 24 were male and 13 were female; no attempt was made to balance males and females across the two conditions. All subjects were native speakers of English, with all but 5 native speakers of a British dialect. The average age was 23 (s.d. 4.6), with computer experience between intermediate/advanced on average, and programming experience between beginner/intermediate.

Subjects were given brief printed instructions describing the usage scenario and suggesting several ways in which they could interact with the system. To motivate subjects to pay attention, they were asked to imagine that they needed to discuss available options with their partner at home, and would be given a chance to take notes on the designs they saw after the interaction. They were also warned that the virtual sales agent would not always understand what they said, and that they could continue by either repeating their request or trying a different one. However, the instructions did not mention that the interaction would be half-duplex, i.e. that the system would not be listening while it was speaking, or after it decided that the user turn was finished; additionally, the subjects were not warned that the system (in its preliminary state of development) would often be rather slow to respond.

After interacting with the system, subjects were given a form which showed pictures of all the available designs, together with the manufacturer and series names. The subjects were asked to write down what they remembered the system telling them about each design that they saw. The recall forms were separately coded by two judges (the first two authors). For each atomic fact conveyed to a subject about a tileset, each judge indicated whether the subject recalled that fact in the notes they took on the recall form; see (White & Foster, 2004) for full details. The two judges then met to discuss the 21 discrepancies between the two codings, and came up with an agreed coding

containing 408 recalled facts across the 37 subjects. Several of the discrepancies involved simple oversights; the remaining ones were resolved by going with the stricter interpretation.

The subjects were also given a questionnaire containing items on 5-point Likert scales, divided into groups for perceived task success, task ease, dialogue quality, intuitiveness, engagement and general liking; four questions eliciting free form comments; and six demographic questions. The questionnaire items drew in part from those listed in (Walker et al., 2000; Sidner et al., 2004); see (White & Foster, 2004) for the complete questionnaire.

In addition to the recall form and questionnaire, we also logged the subjects' interactions with the system. From the logs, we calculated a range of objective metrics as indicators of task success and dialogue quality. In particular, we counted the number of unique tilesets viewed and the number of 3-D tours taken, to use as measures of task success along with recall. Finally, using the videos of the interactions, we counted the number of times the subject looked back from the virtual bathroom display to the avatar.

4 Results

4.1 Objective Metrics

In general, the system worked robustly, with the average dialogue lasting nearly 17 minutes (s.d. 6 minutes), which was longer than we had expected the interaction to remain interesting. Only two subjects experienced technical problems while using the system; we have retained the data from these two subjects in the analysis, as in both cases the bugs appeared after an extended interaction. There were about 25 turns per dialogue, with a large range of 8-57 turns (s.d. over 10). Users viewed more than 9 different tilesets on average (s.d. 3.5, range 2-17); they also took more than 2 3-D tours (s.d. 0.89, range 1-5). We considered these objective measures of task success to be quite promising for this interim version of the system.

The number of turns where the system signalled an error averaged only 1.22, but with substantial variation (s.d. 2.07, range 0-11), and a larger number of time outs, averaging 3.27 (s.d. 2.78, range 0-11). A total of 17% of the turns had either an error message or a time out (s.d. 12%, range 0-50%), indicating that the dialogues did not always go as smoothly as we would have liked. The variation across subjects was especially of concern, as it confirmed our impression that some users had a particularly difficult time using the system. Looking at the relationship between the number of unique tilesets viewed, dialogue length and the error rate, we found that while dialogue length was strongly positively correlated with the number of tilesets viewed ($r=0.70$, $p < 0.001$), the error rate exhibited a fairly strong negative correlation ($r=-0.41$, $p=0.01$, one-tailed), as expected.

Of the user turns, approximately one quarter involved choosing a tileset with a mouse, which was almost always successful. With the system turns, note that the response delay differed substantially between the two face conditions: in the expressive condition, the face offered visual feedback first, in 1.40 seconds on average, while in the zombie condition, there was no feedback before the verbal response, which took 3.99 seconds on average.

Unexpectedly, the subjects in the zombie condition viewed significantly more tilesets than those in the expressive condition (10.67 to 8.26, $p=0.04$, 2-tailed), and had a higher percentage of turns with mouse input (0.27 to 0.23, $p=0.05$, 2-tailed). Finally, as expected from Sidner et al. (2004), we found that the subjects in the expressive condition looked back at the avatar significantly more frequently than those in the zombie condition (2.45/min. to 1.85/min., $p=0.05$, 1-tailed).

4.2 Recall

On average, the subjects were told a total of 39.62 facts and recalled 8.38 of them, for a recall rate of about 21%. We consider this level of recall promising, since the goal of the interaction is not for users to remember everything the system has told them, but rather to retain the information that is of most interest to them. Looking at the relationship between the number of tilesets viewed and recall, we found no correlation ($r=0.02$) with recall rate, while we found a strong correlation ($r=0.49$, $p < 0.001$) with absolute recall, as expected

Interestingly, we found that female subjects had significantly higher scores than male subjects with both absolute recall (11.09 to 6.91, p=0.01, 2-tailed) and recall rate (0.27 to 0.10, p=0.03, 2-tailed). We also found a significant interaction between gender and the face condition, as illustrated in Figure 3 below for absolute recall (relative recall yielded a similar picture); see (White & Foster, 2004) for further details.

It is not obvious what the source of this interaction might be. One possibility is that on average, the males in the study took less interest in the bathroom redesign task, and thus were more easily distracted by what the expressive face was doing. Another possibility is that the females were better able to divide their attention between the bathroom display and the expressive face, and thus much less prone to being distracted from paying attention to the tiling designs. We did not find any other significant interactions between gender and our task success measures or the questionnaire items.

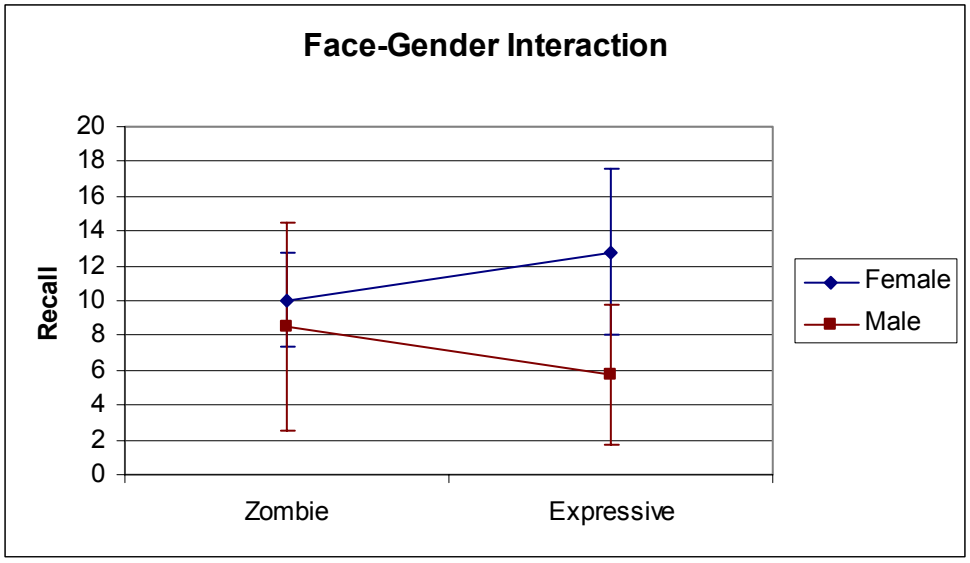

Figure 3: Interaction between Gender and the Face Condition with Absolute Recall

4.3 Questionnaire Items

Overall, perceived task success was good, with an average of almost 4/5 (3.97) for the four items in this category. As expected, absolute recall was positively correlated with perceived success (r=0.41, p=0.01, 1-tailed), as was the number of tilesets viewed, though less so (r=0.24, p=0.08, 1-tailed); the recall rate was correlated at a level between these two (r=0.33, p=0.02, 1-tailed). Conversely, the error rate was strongly negatively correlated with perceived success (r=-0.55, $p < 0.001$).

With the categories related to user satisfaction, we found that overall task ease was slightly positive (3.25), as were quality (3.25) and general liking (3.38); in contrast, overall intuitiveness was neutral (2.94), and overall engagement was slightly negative (2.77). Drilling down, on the positive side, we found that subjects tended to *agree* that

- it was easy to look at a range of tiling designs (Q6, somewhat) and to take a 3-D tour (Q8);
- the system understood what they pointed to (Q10); the system's pointing gestures were natural (Q12) and helpful (Q13); the voice was easy to understand (Q16); and the system gave useful information (Q17) which was easy to follow (Q18);
- the system worked the way they expected it to (Q21, somewhat) and was not complicated (Q31);
- they did not have to concentrate too hard to use the system (Q36); and
- they liked the system (Q38, somewhat), found it friendly (Q39) and knowledgeable (Q40); and thought it gave them accurate information about the tiles (Q42).

However, on the negative side, we also found that subjects tended to *disagree* that

- the system understood what they said (Q9) and responded quickly (Q11); the facial expressions seemed natural (Q14, somewhat) and helpful (Q15, somewhat); and the system conveyed the right amount of information at once (Q19, somewhat) without being repetitive (Q20);
- they knew what they could say or do at each point (Q23) and when to begin speaking (Q24); the system was flexible (Q27, somewhat); and they felt in control (Q28); and
- the conversation was engaging (Q32), involving (Q34), and not boring (Q37, somewhat).

The cases where t-tests revealed a significant difference between the face conditions appear in Table 2 below:

Table 2: Questionnaire Items with Significant Differences between Face Conditions

	Expressive		*Zombie*			
Metric / Item	*Mean*	*S.D.*	*Mean*	*S.D.*	*p*	*tails*
Overall Success (Q1-Q4)	3.77	0.67	4.18	0.40	0.031	2
Q5. It was easy to use the system.	2.68	0.95	3.33	1.03	0.053	2
Q6. It was easy to look at a range of tiling designs.	2.95	1.03	3.61	0.92	0.046	2
Overall Ease (Q5-Q8)	2.95	0.75	3.57	0.77	0.017	2
Q15. I found the facial expressions helpful.	**3.05**	1.13	**2.28**	0.89	**0.014**	1
Q28. I felt in control when using the system.	2.42	0.84	3.00	0.91	0.051	2

In line with our prediction that facial expressions would improve usability, the average response for questionnaire item 15, concerning the helpfulness of these expressions, was significantly higher in the expressive condition. However, contrary to our expectations, the subjects in the zombie condition gave higher average responses for overall perceived success (average of items 1-4) and overall ease (average of items 5-8); and with borderline significance, they also gave higher average responses for items 5, 6 and 28, concerning how in control subjects felt.

4.4 User Comments

The subjects provided a large number of comments (see White & Foster, 2004, for the complete list); examples of the most frequent ones appear in Table 3:

Table 3: Example User Comments

Category	*Comments*	*Example*
ASR problems	32	It didn't understand me. Then I got distracted trying to make it understand.
Mouse easy	22	Using the mouse [was the easiest part] - the avatar always understood what I meant.
Slow response	22	I wanted to move around the designs and displays quickly and I didn't feel the system allowed me to do this.
Not knowing when to speak	14	Knowing when to speak so that the system was taking in what you said [was the hardest part].
Face unnatural	13	[A] more friendly looking face would be better, including hair.
Better content	12	The comments the head makes should sometimes be a bit more personal rather than sounding like he is reading the manufacturer's brochure. More descriptive words like fresh, airy, easy to clean etc.
Not knowing what can be done	12	Perhaps [the system could be improved] by have a small options bar at the bottom. It could have designs you would like to see again […] and perhaps have a list of the main types of tiles ("classic", "modern" etc) so people can remember what's available.
No barge-in	10	It has to understand faster. It should also be okay with me interrupting by clicking or talking.
Use GUI	8	The system cannot do anything that could not be done quicker with a mouse. It just has a set of commands it can follow which might as well be buttons on a GUI.
Voice	7	Some of the speech sounded fragmented or pieced together - sometimes like it was

Category	*Comments*	*Example*
problems		starting to say something and finished with something else.
Not knowing what to say	7	Could the sys. be improved by a help menu for the allowed dialog to use?

The counts of user comments—and subjects making those comments—appear in Table 4 below, by category and condition (X for expressive, Z for zombie), and sorted by frequency. Note that some subjects made comments falling into the same category multiple times, so the subject counts are sometimes lower than the comment counts.

Table 4: Number of User Comments by Category and Face Condition

	Comments			*Subjects*		
Category	*Total*	*X*	*Z*	*Total*	*X*	*Z*
ASR problems	32	18	14	22	12	10
Mouse easy	22	9	13	22	9	13
Slow response	22	7	15	15	**4**	**11**
Not knowing when to speak	14	7	7	13	6	7
Face unnatural	13	4	9	12	4	8
Better content	12	5	7	12	5	7
Not knowing what can be done	12	7	5	9	5	4
No barge-in	10	3	7	8	3	5
Use GUI	8	6	2	8	6	2
Voice problems	7	4	3	7	4	3
Not knowing what to say	7	5	2	5	3	2
3-D tours easy	6	3	3	6	3	3
Voice good	6	2	4	5	2	3
Didn't look at face	4	1	3	4	1	3
Face good	4	2	2	4	2	2
Better feedback	4	1	3	3	1	2
Yes-no questions easy	3	2	1	3	2	1
Practice	3	1	2	3	1	2
Knowing what to do easy	2	1	1	2	1	1
ASR good	2	1	1	2	1	1
Crash	2	2	0	2	2	0
Go back	1	0	1	1	0	1
Other	18	11	7	13	7	6

Applying a chi-square test of independence revealed a significant difference in the *slow response* category. The observed frequencies for the number of subjects mentioning slow response is given in the Table 5 below, followed by the frequencies expected under the null hypothesis that mentioning slow response is independent of the face condition.

Table 5: Observed vs. Expected Frequencies of Mentioning Slow Response by Face Condition

OBSERVED FREQUENCIES	*Mentioned slow response*	*Didn't mention it*	*Row Total*
Expressive condition	4	15	19
Zombie condition	11	7	18
Total	15	22	37

EXPECTED FREQUENCIES	*Mentioned slow response*	*Didn't mention it*	*Row Total*
Expressive condition	7.70	11.30	19
Zombie condition	7.30	10.70	18
Total	15	22	37

According to the chi-square test, the probability that just 4 out of 15 subjects mentioning slow response would be in the expressive condition is 0.01. This suggests that the visual feedback provided by the facial expressions in advance of the verbal responses helped to mitigate the system's perceived sluggishness in responding.

5 Discussion and Future Research

Our hypothesis that the avatar's facial expressions would improve usability was partially confirmed. In particular, we found significant differences between the two conditions on one questionnaire item concerning whether the avatar's expressions were helpful, and on the number of comments mentioning the system's slow response. Together, these differences indicated that the thinking expression helped to convey that the system was busy processing input, confirming Edlund and Nordstrand's (2002) finding, and that the subsequent expressions mitigated the system's perceived sluggishness in responding verbally.

We had also hoped to show that the avatar's looking behaviour would improve conversational efficiency by helping to signal to the user when the system had come to the end of its turn and was thus ready for input. Unfortunately though, technical problems prevented us from investigating this hypothesis fully, as the avatar often ended up turning back towards the user in the middle of the system turn, rather than before the final system utterance. The looking behaviour may nevertheless have played a role in the surprising negative effects of the expressive face condition on task success, task ease and feeling in control. In Sidner et al.'s (2004) study involving technology demonstrations given by a talking robotic penguin, they found that in the condition where the robot could turn its head towards the demonstration table, subjects paid more attention to the robot, and appear to have adjusted their looking based on the robot's looking. In line with their findings, we observed in the videos that the subjects in the expressive condition looked back at the avatar significantly more often than those in the zombie condition. However, in contrast to Sidner et. al's study, we found that the expressive face had a negative impact on task success and ease. This may have been because the expressive face to some extent distracted subjects from the task of examining the different tiling possibilities, without improving the interaction enough to compensate for the distraction. Another possibility is that the expressive face raised users' expectations of the system's abilities, thereby encouraging subjects to use voice input rather than the mouse relatively more often than in the zombie condition, which was generally a less successful strategy.

Finally, we had expected that the avatar's expressions would make the interaction seem more natural and thus improve user satisfaction. However, we found no significant differences in the perceived naturalness of the face, general liking, or overall satisfaction. Note that twice as many subjects in the zombie condition mentioned that the face was unnatural than in the expressive condition, but the frequencies were too low to reach significance. Possible reasons that general liking and overall satisfaction were not significantly affected may have been that the positive and negative effects of the expressive face cancelled each other out, or that any effect of the face condition was swamped by the overall clumsiness of the multimodal interaction.

In hindsight—and after examining the videos of the user interactions—we concluded that while the avatar feedback was helpful with some users, it should not be expected to make up for the unnaturalness of a half-duplex interaction on its own. For example, while the avatar was displaying its thinking expression, several users tried repeating their requests on multiple occasions—presumably in the hope that the repeated request would be easier for the system to understand, unaware that the microphone was actually turned off at this point. Interestingly, this conclusion is supported by de Ruiter's (2005) recent research on human-human dialogues involving relevant external visual representations (like the bathroom display in the COMIC system), where he found no systematic relationship between other-gaze and turn-taking in human-human dialogues: the conversational participants were able to converse smoothly with very little other-gaze, which played no apparent role in exchanging the floor. Of course, this is not to say that looking behaviour cannot play a useful role in dialogue, for example to direct attention (cf. Sidner et al., 2004), or to modulate the flow of information (cf. Nakano et al., 2003).

With this conclusion in mind, in the final version of the COMIC system—which remains half-duplex—we have augmented the avatar's facial feedback with explicit signals of when the system is prepared to accept input, including busy cursors and disabled widgets. In response to user feedback, we have also incorporated more options for mouse as well as voice input, improved system response times, and reduced the average length of system turns.

Given this improved interaction design, we plan to investigate further whether an expressive avatar can improve naturalness and user satisfaction, without compromising task performance.

Acknowledgements

Thanks to Holly Branigan and Guy Whitten for advice on designing the experiment and interpreting the results, and to Louis ten Bosch, Lou Boves, John Lee, Johanna Moore, Els den Os, and Jan Peter de Ruiter for helpful discussion. This work was supported in part by the COMIC project, IST-2001-32311.

References

Cassell, J., & Thorisson, K. R. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13, 519–538.

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., & Yan, H. (2001). More than just a pretty face: conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*,14, 55–64.

Edlund, J., & Nordstrand, M. (2002). Turn-taking Gestures and Hour-Glasses in a Multi-modal Dialogue System. In *Proc. ISCA Workshop Multi-Modal Dialogue in Mobile Environments*.

Foster, M. E., & White, M. (2004). Techniques for Text Planning with XSLT. *Fourth NLPXML Workshop*

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, 26, 22–63.

Nakano, Y., Reinstein, G., Stocky, T., & Cassell, J. (2003). Towards a model of face-to-face grounding. In *Proc. ACL-03*.

den Os, E., & Boves, L. (2004). Natural multimodal interaction for design applications. In *Proc. eChallenges e2004*.

de Ruiter, J. P. (2005). The role of other-gaze in dialogue involving relevant external visual representations. In de Ruiter, J. P., Schmiedtová, B., & Chen, A., eds., *Research on modality effects on performance quality and efficiency*, COMIC project deliverable 2.3, available from http://www.hcrc.ed.ac.uk/comic/documents/.

Sidner, C. L., Kidd, C., Lee, C., & Lesh, N. (2004). Where to look: A study of human-robot engagement. In *Proc. ACM International Conference on Intelligent User Interfaces (IUI)*, pp. 78–84.

Walker, M. A., Kamm, C. A., & Litman, D. J. (2000). Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*.

White, M., & Foster, M. E. (2004). Report on Human Factors Experiments with the Integrated T24 Demonstrator (part 2). COMIC project deliverable 1.3b, available from http://www.hcrc.ed.ac.uk/comic/documents/.